A Survey on Multimodal Learning Methods in the Context of Speech

Shikhar Gupta
shikhar.gupta@utexas.edu

Geethika Hemkumar
geethika.hemkumar@utexas.edu

Abstract

Multimodal learning involves the processing and synthesis of multiple data modes. In this survey paper, we focus on multimodal learning involving speech. Two main findings are discussed: the ability for multimodal learning to improve classical speech tasks, and new tasks/processing capabilities introduced by multimodal learning. We additionally discuss the different styles of multimodal fusion in the context of the papers surveyed. Finally, we discuss the multimodal chain method, which utilizes multimodal inputs in a different manner than other methods we surveyed. As a result of our survey within this area, we believe that multimodal learning is a topic worth continued study due to the novel reasoning capabilities and learning tasks it enables.

1 Introduction

Multimodal learning involves the processing and synthesis of multiple *modes* of data inputs. A mode is a method of information transmission, such as: text, video, audio, speech, or image. Multimodal learning has been recently utilized in several areas, such as robotics [14], vision [1], and speech, which we'll focus on in this paper.

One motivation for further research in multimodal learning is the the McGurk Effect, a sensory illusion that arises due to a conflict in audio and visual perception. The results of the original study [17] showed that simultaneously presenting visual data of a speaker's lips articulating an utterance different from that sounded in an audio clip affected identification of the speech utterance that participants heard. Without presentation of the visual component, the study participants reported hearing the correct utterance with high accuracy. If both the audio and visual data are fed into a multimodal model, then the model can learn to recognize and address any discrepancies (i.e., in transcription) due to the McGurk Effect.

In this report, we present a survey on multimodal learning methods involving speech. In Section 2, we synthesize our findings from our literature survey. Two main themes are discussed: the potential for improvement in performance of classical speech tasks as a result of multimodal learning, and the introduction of new model capabilities by multimodal learning. An important aspect of multimodal learning is how the model handles multimodal fusion; Section 3 elaborates on the various approaches. In Section 4, we delve into an alternate application of multimodal inputs, the multimodal *chain* approach. We discuss challenges and future work in Section 5 and we conclude by summarizing the important findings of our survey.

2 Findings and Trends

We organize the findings from our literature survey into two categories. Firstly, we discuss how multimodal learning can improve classical speech tasks, which are often originally unimodal. Secondly, we discuss new learning tasks/capabilities that are introduced by multimodal learning.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

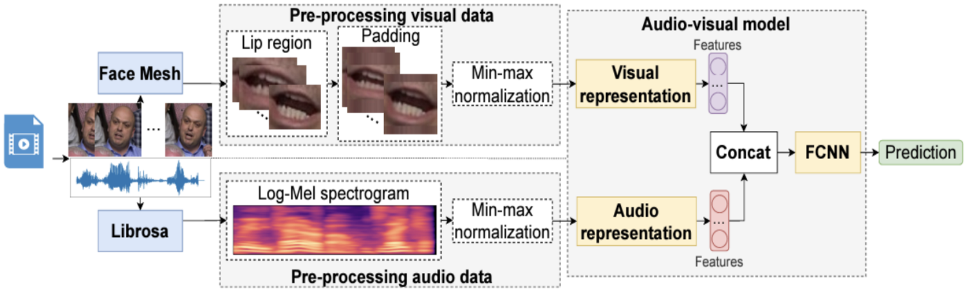


Figure 1: The architecture for the Audio-Visual ASR approach taken by [23].

2.1 Improving classical speech tasks with multimodal learning

Introducing multiple input modalities into architectures can provably improve the performance on classical speech tasks. In this section, we'll focus specifically on multimodal architectures, specifically those that include video and image inputs, that enable improvements on the ASR task. This particular subtask is called *Audiovisual* ASR (AV-ASR). We will look at two approaches to this, one based on a concatenation of individual processed features, and one based on a cross-attention mechanism.

2.1.1 Enhancing ASR Through Lip Reading

Ryumin et al. [23] introduce a new multimodal architecture which accepts both video and audio data to produce a prediction on the ASR task. The intuition behind this approach is based on the concept of *inverse effectiveness*, which poses that human speech comprehension is improved by being able to see a speaker's face, particularly in noisy conditions with many speakers [6]. In particular, the authors make use of the Lip Reading in the Wild (LRW) dataset [5]. This dataset consists of short utterances (1-2s each) from BBC news segments and talk shows. There are more than 1000 different speakers and 500 recognition classes (words).

The architecture is shown in Figure 1. The visual and audio information is preprocessed through separate pipelines, and the final feature vectors are concatenated and passed through a fully connected network to produce a prediction. For the visual portion, the authors focused particularly on the lip regions of the images, extracting these features through the MediaPipe Face Mesh [10] library and normalizing them using min-max normalization. After this, the final feature vectors were extracted via a variant of the ResNet-18 [12] architecture. For the audio portion, the Librosa [16] library was used to preprocess the audio signal into log-Mel spectrogram features and also normalized using min-max normalization. The final feature vectors were extracted, similar to the visual features, via a variant of the ResNet-18 [12] architecture. Though the architectures for processing audio and visual features were relatively similar, the authors processed them through separate pipelines before fusion occurred, and this yielded better results than other methods. This is discussed further in Section 3.

Through this architecture, the model is able to achieve *98.76%* recognition accuracy, which is the highest achieved accuracy when compared to state-of-the-art results on the LRW dataset. Some further study that can be done with this approach could be testing against other, purely unimodal ASR models to test for the difference in WER when feeding in audio-only versus audio + visual data, or training the model with the entire face image rather than by only extracting lip region information to see if facial expressions could add valuable information. However, these results are promising and definitely indicate the benefits of including visual information to enhance the performance of ASR.

2.1.2 AV-ASR with Cross-Attention

Paraskevopoulos et al. [20] present a Transformer-based architecture for performing ASR by incorporating visual and audio modalities. The intuition behind using a transformer-based architecture is to provide context and grounding to the auditory input modality through the video input. For example, a clip from a game show would have a different probability distribution for words than one from a

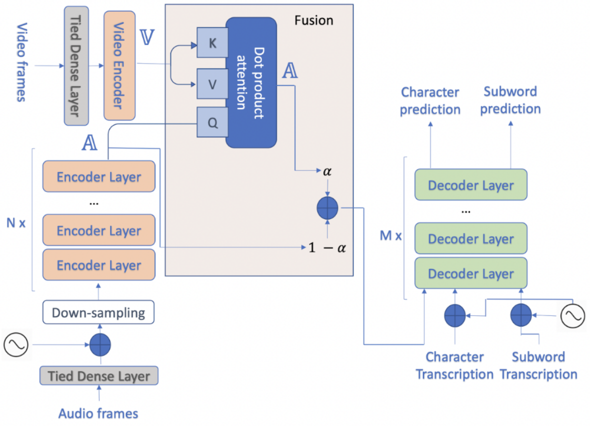


Figure 2: The architecture for AV-ASR with Cross-Attention as utilize in [20].

Features	Level	WER	⇑ over audio
Audio	Subword	26.1	-
Audio + ResNeXt	Subword	25.0	3.45%
Audio	MR	**21.3**	-
Audio + ResNeXt	MR	20.5	3.76%
Audio (B)	Subword	**19.2**	-
Audio + ResNext (B)	Subword	**18.4**	3.13%

Figure 3: Comparison of audio-only versus audio-visual models on the ASR task for [20]. (B) indicates the results from the LAS architecture [4].

baseball game. Additionally, this approach was motivated by the positive results achieved in NLP tasks by using transformer-based architectures [28] [7].

The architecture is shown in [20]. The audio and visual input modalities are first preprocessed through a dense layer and encoder layer(s) to produce preliminary features. Then, the dot product cross-attention layer is used to project the visual features onto the audio feature space before the combined output features are passed through decoder and dense layers to produce the final predictions. The main contribution of this architecture is the cross-attention layer, which effectively fuses the information from the audio and visual modalities, and serves to contextualize the audio information received through the signal.

The results 3 also support the claim that the video information contributes towards the overall efficacy of the AV-ASR system. The authors compared the WER on the How2 [24] dataset, which consists of over 300 hours of instructional video data aligned at the word-level with English transcriptions, sourced from YouTube. The authors compared the ASR performance from three audio-only models, including the Listen, Attend, and Spell audio-only architecture [4], on which adding image features actually improved the performance. Across all models, >3% improvement performance was observed when adding visual features to the architecture, indicating the benefits of the multimodal approach once again.

2.2 Enabling new model capabilities

Multimodal learning can also introduce unique model capabilities. In this section, we discuss how multimodal learning can enable a single model to perform multiple speech tasks, elaborate on tasks that multimodal learning is uniquely capable of achieving, and discuss model architecture choices that enable flexible inclusion/exclusion of multimodal inputs to the model.

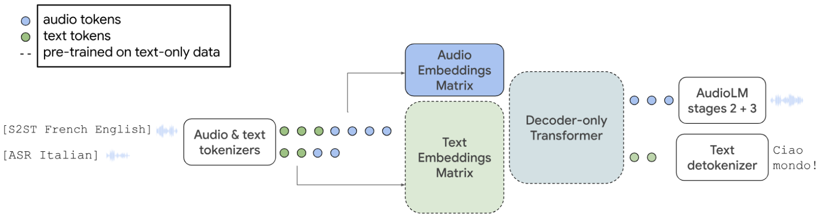


Figure 4: The AudioPaLM architecture as shown in [22]

Model	CoVoST2 AST BLEU↑	CVSS S2ST ASR-BLEU↑	VoxPopuli ASR WER↓
Whisper Large-v2 1.5B [Radford et al., 2022]	29.1	–	13.6
mSLAM-CTC 2B [Bapna et al., 2022]	25.2	–	9.1
MAESTRO 600M [Chen et al., 2022c]	25.2	–	**8.1**
USM-M [Zhang et al., 2023a]	30.7	–	–
Translatotron 2 + pretraining + TTS aug [Jia et al., 2022a]	–	25.6	–
AudioPaLM 8B AST (ours)	35.4	–	11.1
AudioPaLM 8B S2ST (ours)	36.2	**32.5**	16.0
AudioPaLM-2 8B AST (ours)	**37.8**	–	9.8
AudioPaLM-2 8B cascaded ASR + transl. (ours)	39.0	–	–

Figure 5: AudioPaLM experimental evaluations from [22] on ASR, AST, and S2ST compared to baselines

2.2.1 Enabling Performance of Multiple Tasks with One Model: AudioPaLM

AudioPaLM [22] is a speech model that learns from both text and audio data. This model is a fusion of PaLM-2 [2], a text-only model, and AudioLM [3], an audio-only model. By combining these single-modal models into a single multimodal model, AudioPaLM can utilize both the language reasoning capabilities from PaLM-2 and the audio understanding capabilities, such as preserving a speaker's voice in audio generation, from AudioLM.

Figure 4 shows the architecture of the AudioPaLM model. First, the text and audio inputs are tokenized separately. Using the tokenized input, a combined text-audio embeddings matrix is created. A decoder-only Transformer models both text and audio based on this embeddings matrix and generates either text or audio tokens depending on the specified task. Finally, depending on the specified task being performed, either the SentencePiece text detokenizer or portions of the AudioLM model (stages 2 and 3) is used to decode the tokens generated from the Transformer into raw text or audio.

By learning a joint speech and text representation, AudioPaLM is capable of performing a variety of single-modal speech tasks with a single model. These tasks include: ASR, AST, MT, TTS, and S2ST. In this case, multimodal learning expands the versatility of a single model.

Figure 5 shows the results from [22] of evaluating AudioPaLM on ASR, AST, and S2ST tasks. AudioPaLM's performance on S2ST and AST exceeds those of the baselines, and achieves strong performance compared to the best-performing ASR model. Importantly, AudioPaLM is able to achieve state-of-the art performance on several tasks using a single model. Having a single model that can perform several tasks effectively is attractive for real-world applications, as a single model can be trained to perform multiple tasks that previously would have required training several separate models.

Another set of results highlighted in the AudioPaLM paper is model training on combined tasks, where a given task is broken down into a series of intermediate tasks. Further, the result from the previous intermediate task is used as input to the next task. Figure 6 shows the results from [22] of additionally training AudioPaLM on direct (i.e., no intermediate tasks) and combined S2ST tasks. Though the model performance on AST and ASR tasks is slightly worse on this model compared to the model that can only perform AST and ASR, the model gains the ability to perform S2ST tasks.

Tasks	CoVoST2 AST	CVSS S2ST	CoVoST2 ASR
	BLEU↑	ASR-BLEU↑	WER↓
AST, ASR	30.5	–	25.3
AST, ASR & S2ST	27.8	24.2	27.1

Figure 6: Results of adding S2ST capabilities to AudioPaLM from [22]

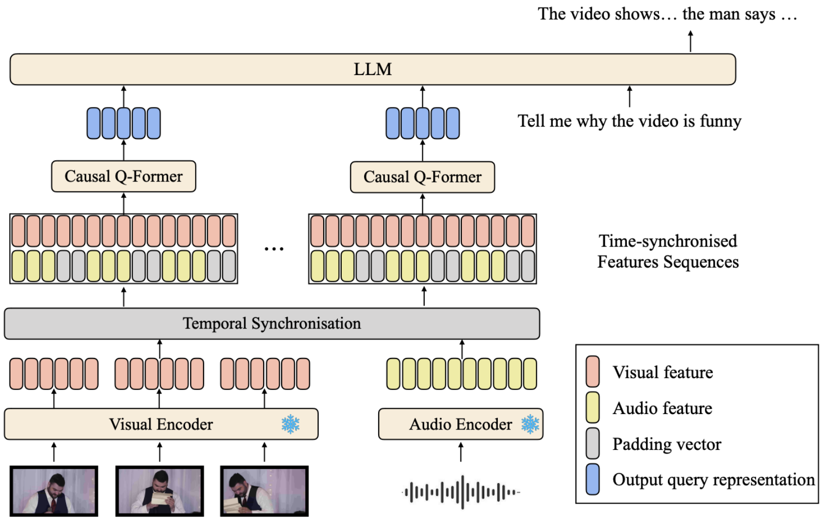


Figure 7: The FAVOR architecture, from [26]

2.2.2 Introducing New Tasks: FAVOR

FAVOR [26], short for *fine-grained audio-visual joint representation*, is an audio-visual learning model. By learning a joint audio-visual representation, FAVOR is able to synthesize information from both the audio (speech and other forms of sound) and visual (images/videos) perspectives when reasoning about a video.

Figure 7 shows the model architecture. The temporal synchronization module aligns the audio and visual data streams. This time-aligned representation is passed to the causal Q-Former layer, which is designed to realize correspondences between the synchronized audio and visual streams of data. More specifically, a causal self-attention module is added within the Q-Former layer to focus on extracting information based on the sequential nature of both the audio and visual data (i.e., using information from prior frames of audio/video to explain the current ones).

This joint audio-visual understanding enables FAVOR to perform new audio-visual tasks such as: image spoken question answering (ISQA) and audio-visual matching (AVM). The former task involves using an image/images to answer questions given through audio, and the latter involves determining whether or not the information in the audio and visual streams match. Such tasks cannot be performed by a model that only understands visual or audio inputs.

Figure 8 shows the results from [26] of evaluation of FAVOR on a variety of audio-visual tasks. FAVOR's performance exceeds those of prior models, demonstrating the strength of its joint audio-visual reasoning capabilities.

The key advantage of audio-visual models such as FAVOR is its ability to provide context about a video (containing audio) from a perspective that unimodal models cannot.

Systems	AVSR ↓	AVSD ↑	ISQA ↑	AVSSD ↑	AVM ↑
Whisper large-v2	8.3%	-	-	-	-
InstructBLIP 13B	-	41.4%	-	1.1%	-
InstructBLIP 13B fine-tuned	-	52.1%	-	20.3%	-
Video-LLaMA 7B	-	27.6%	-	41.9%	52.3%
FAVOR 13B (ours, audio-only)	8.3%	-	-	34.7%	-
FAVOR 13B (ours, visual-only)	-	53.3%	-	23.5%	-
FAVOR 7B (ours, audio-visual)	8.7%	51.2%	24.5%	50.5%	74.3%
FAVOR 13B (ours, audio-visual)	**8.1%**	**54.5%**	**32.3%**	**51.1%**	**77.1%**

Figure 8: The results of evaluation of FAVOR on audio-visual tasks, from [26]. AVSR is audio-visual speech recognition, AVSD is audio-visual scene-aware dialogue, ISQA is image spoken question answering, AVSSD is audio-visual sound source detection, and AVM is audio-visual matching.

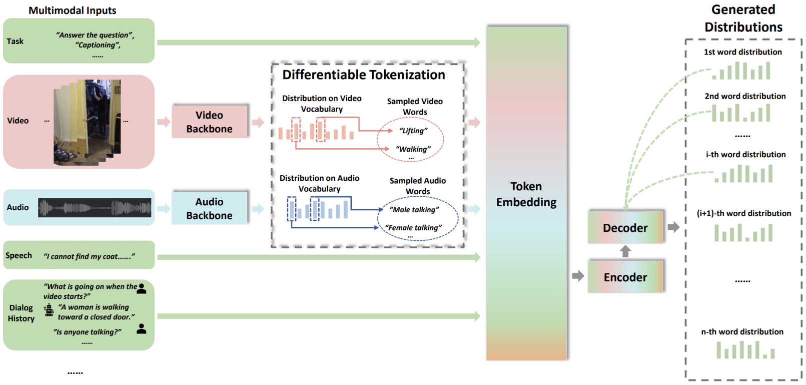


Figure 9: The VX2Text model architecture, from [15]

2.3 Expanding Multimodality: VX2Text

VX2Text [15] is a multimodal model for text generation from a combination of video + "x" inputs, where "x" is any modality such as text, speech, or other forms of audio. VX2Text's model architecture enables the fusion of video and combinations of inputs of any other modality.

Figure 9 provides an overview of the model architecture. A modality-specific classifier encodes raw data of this modality into a text representation. Embeddings of these results are fused via a transformed encoder/decoder, which produces text summarizing the information contained in all of the original inputs.

Performing fusion through language enables VX2Text to incorporate data of any input modality into text generation for a given task given the existence of a classifier that maps the input data to a text representation.

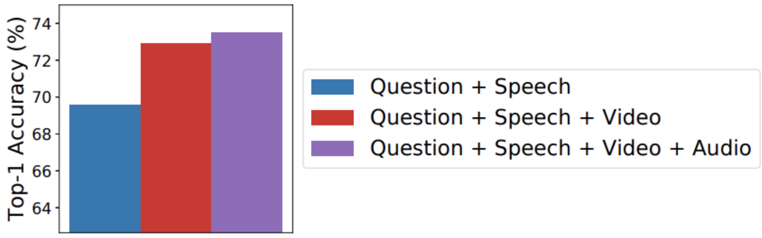


Figure 10: Top-1 Accuracy of VX2Text's performance on the VQA task using the TVQA dataset given various combinations of input modalities, from [15]

One experiment within the VX2Text paper evaluates the benefit of each input modality to the performance of the video question answering (VQA) task on the TVQA dataset, where the model must respond to a question about a TV video clip and its transcribed speech. Figure 10 shows the results of this experiment, which demonstrates the incremental gain in top-1 accuracy with the addition of each input modality. These results convey the ability for VX2Text to extract additional information from each modality that could not be provided by the other modalities.

3 Multimodal Fusion Methods

One of the most important considerations when designing a multimodal architecture is the method of combining input modalities and the stage within the architecture where a combined representation is created. We generally section the various approaches into three distinct categories: (1) *early, or feature-level fusion*, in which a combined representation is created at the data level or very early in the architecture before major processing or learning has occurred, (2) *intermediate, or model-level* fusion, where a combined representation is created after major preprocessing, feature extraction, or learning has occurred but before any output, and (3) *late, or prediction-level* fusion, where input modalities are processed separately and individual outputs are combined to produce a final prediction. We explore the considerations, benefits, and potential drawbacks of each.

3.1 Early Fusion

Early fusion occurs when data from input modalities are fused directly or before any major learning/feature processing is performed. Compared to other forms of fusion, early fusion approaches minimize preprocessing of individual modalities before fusion occurs and a combined representation is created, which may allow the model to capture more information about dependencies between modalities than other approaches. However, because there is little to no preprocessing on each input modality prior to fusion, noise from individual inputs can cascade in the combined representation, and interpretability of results is occluded, making the model difficult to fine-tune.

In the AudioPaLM [22] architecture, the tokenized text and audio are concatenated into a single embeddings matrix before any major feature processing is performed. Though fusion occurs after tokenization, the tokenizers are pre-trained models that are not updated during AudioPaLM model training. The majority of learning thus occurs in the decoder-only Transformer, which is after modality fusion occurs.

A similar argument can be applied to the FAVOR [26] architecture. The visual and audio encoders are not updated during FAVOR's model training, and the temporal synchronization module is not learned. Hence, the majority of learning occurs in the causal Q-Former layers, which receive the fused audio-visual representation as input. As previously discussed, the early fusion approach combines inputs at or near the data level, which may yield better learning of interdependencies between inputs compared to other approaches. The early fusion approach thus makes sense in this scenario, since FAVOR aims to learn a *fine-grained* audio-visual representation.

3.2 Intermediate Fusion

Intermediate, or model-level multimodal fusion of modalities occurs after preprocessing and feature extraction is performed on the input modalities separately. The intuition behind intermediate fusion is to extract the "useful" features from each input modality before fusion occurs to maximize the expressivity of the combined representation as much as possible. After fusion occurs, the combined representation of input modalities is then passed through one or more additional neural network model layers before a final prediction or output is produced. This method of fusion is the broadest, as fusion can generally occur at any point in the architecture after feature extraction and before output and still be considered to be intermediate fusion.

Ryumin et al. [23] make use of intermediate fusion in fusion audio and visual input modalities for the ASR task. Both input modalities are processed by a variant of the ResNet-18 [12] architecture to extract feature vectors before fusion occurs. Here, the method of fusion is a simple concatenation of the extracted feature vectors. The bulk of the processing actually occurs before the modalities are combined, in fact, since once the combined representation is created, the prediction is produced by a simple fully connected network. We would consider this to be on the later end of intermediate fusion, therefore.

Paraskevopoulos et al. [20] also utilize intermediate fusion for audio and visual input modalities for the ASR task. However, the architecture is on the earlier side of intermediate fusion since audio and visual input modalities are passed only through encoder layers before fusion occurs. The authors also utilize a novel approach to fusion in this architecture with a cross-attention mechanism that projects the visual features onto the audio space, in contrast to the concatenation approach generally utilized for fusion in most multimodal architectures. This not only creates a combined representation but the cross-attention layer also explicitly learns relationships between the input modalities, which can create a more effective model, as evidenced by the positive results 3.

VX2Text [15] is also an intermediate fusion model, as a separate classifier per input modality is used to convert the raw data into a language representation. Fusion occurs within a Transformer encoder-decoder. This occurs after modality-specific processing is completed and before a prediction, the generated text, is produced.

Thus, the intermediate fusion approach provides flexibility in multimodal architectures for the fusion method and placement within the architecture. This approach can provide a good middle-ground for maximizing expressivity from individual features while still being able to learn from a combined representation.

3.3 Late Fusion

Late, or prediction-level fusion, occurs right before the final prediction or output of a multimodal architecture. Generally, the input modalities are processed separately, and outputs are also produced separately with respect to each modality, without any interaction between modalities. Then, the predictions themselves are combined in a meaningful way to produce a final output. Late fusion is generally analagous to running multiple models with the same input, and it is the least truly multimodal in nature than the previous methods of fusion. It is also the least common method for multimodal fusion. However, late fusion allows for the most flexibility since model types, architectures, and hyperparameters can be adjusted for any input modality without affecting the performance of the others.

Schuller et al. [25], utilized a late-fusion model for the task of speech emotion recognition, a notoriously difficult task. The authors utilized five different speech-based emotion recognition engines (some based on auditory input, and some based on textual input), and used a "democratic" voting system based on the final individual predictions to determine the final output. This late-fusion approach based on voting outperformed the best individual classifier.

Ryumin et al. [23], while their main multimodal architecture utilized an intermediate fusion approach, also experimented with a prediction-level approach. In this approach, the video and auditory predictions are combined by in a weighted manner by using a Dirichlet distribution to produce a tensor of 1000 x 500 x 2, where there are 1000 randomly generated 500 x 2 matrices, 500 being the number of classes, and 2 being the number of models. The matrix with the best performance on the validation set is used for the final vector that determines the output. This model achieved a 96.87% recognition accuracy, slightly lower than the 98.76% achieved by the intermediate-fusion model, but still a comparable result.

Though late-fusion is a less common approach and may have a limited utility in multimodal learning as compared to early or intermediate fusion approaches, it is much simpler to implement as it involves only combining final prediction outputs. Additionally, it can still produce favorable outcomes close to or better than comparable state of the art unimodal models [23] [25].

4 Alternative Application of Multimodal Learning: Multimodal Chain

Thus far, we have discussed multimodal learning architectures which generally include creating and learning from combined representations of various input modalities, with fusion at various stages within the multimodal architecture. In this section, we explore an alternative approach to multimodal learning, called the *multimodal chain*, proposed by Effendi et al. [8].

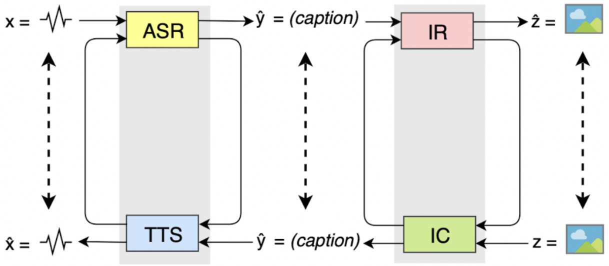
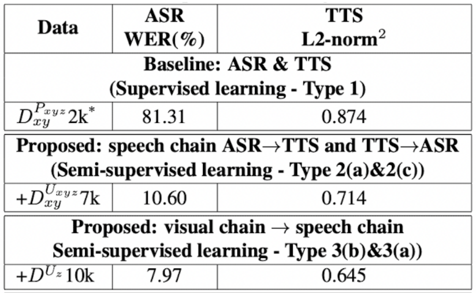


Figure 11: The Multimodal Chain architecture, as proposed by [8].

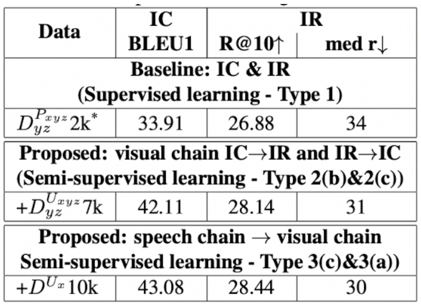
In this approach, a "chain" architecture combines ASR (speech → text), TTS (text → speech), Image Captioning (IC) (image → text), and Image Retrieval (IR) (text → image) models, as shown in Figure 11. Though three input modalities are required to propagate through the entire chain, each individual model only requires one input modality. This allows the training of each individual component without requiring fully paired data, and a self-supervised approach with unpaired data can also be utilized to improve the models. In particular, the authors define three settings:

1. **Fully paired data**: In this setting, a fully supervised learning approach can be taken since there is an image, text, and audio paired transcription for each data point. This is the ideal case, but gathering a large amount of paired data is difficult in practice.
2. **Unpaired speech, text, and image data exist**: In this setting, a self-supervised approach can be used by utilizing reconstruction loss. For example, using only speech data x, ASR generates a transcription $\hat{y}$ which can then be used to reconstruct $\hat{x}$ via TTS.
3. **Single data (data for only 1 modality exists)**: In this setting, we can perform self-supervised learning and unsupervised learning. For example, if only speech data exists, for speech data x, ASR could generate transcription $\hat{y}$ which can then be used to separately perform TTS (self-supervised learning) and IR → IC as well in the visual chain (unsupervised learning).



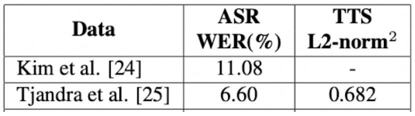
Data	**ASR WER(%)**	**TTS L2-norm^2**
Baseline: ASR & TTS (Supervised learning - Type 1)		
$D_{xy}^{P_{xyz}}$ 2k*	81.31	0.874
Proposed: speech chain ASR→TTS and TTS→ASR (Semi-supervised learning - Type 2(a)&2(c))		
$+D_{xy}^{U_{xyz}}$ 7k	10.60	0.714
Proposed: visual chain → speech chain Semi-supervised learning - Type 3(b)&3(a))		
$+D^{U_z}$ 10k	7.97	0.645

(a) ASR and TTS performance using Multimodal Chain



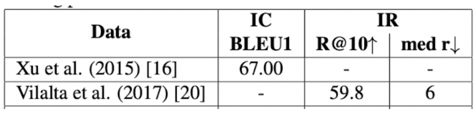
Data	**IC**	**IR**	
	BLEU1	**R@10↑**	**med r↓**
Baseline: IC & IR (Supervised learning - Type 1)			
$D_{yz}^{P_{xyz}}$ 2k*	33.91	26.88	34
Proposed: visual chain IC→IR and IR→IC (Semi-supervised learning - Type 2(b)&2(c))			
$+D_{yz}^{U_{xyz}}$ 7k	42.11	28.14	31
Proposed: speech chain → visual chain Semi-supervised learning - Type 3(c)&3(a))			
$+D^{U_x}$ 10k	43.08	28.44	30

(b) IC and IR performance using Multimodal Chain



Data	**ASR WER(%)**	**TTS L2-norm^2**
Kim et al. [24]	11.08	-
Tjandra et al. [25]	6.60	0.682

(c) ASR and TTS State of the Art Performance



Data	**IC**	**IR**	
	BLEU1	**R@10↑**	**med r↓**
Xu et al. (2015) [16]	67.00	-	-
Vilalta et al. (2017) [20]	-	59.8	6

(d) IC and IR State of the Art Performance

Figure 12: Performance on the four tasks of the Multimodal Chain architecture utilizing a small, paired dataset, then self and unsupervised approaches 12a12b and the associated state of the art results 12c 12d.

The authors also found that applying self-supervised and unsupervised learning approaches in an iterative fashion when only a small amount of paired data is available can still yield favorable results comparable to state of the art results. For all 4 tasks, the authors trained an initial model using 2k paired data points, then trained on a further 7k unpaired data points related to the task at hand (for

ASR/TTS this was unpaired speech and text data and for IC/IR this was unpaired image and text data). Lastly, the models were trained on a further 10k single data-type points for unsupervised learning (for ASR/TTS this was image data used to train the ASR/TTS task and for IC/IR this was audio data used to train the IC/IR task). Through this iterative process, the authors were able to achieve results comparable to existing state of the art methods [13][27][29] [9]for all 4 tasks. Perhaps the most interesting result is the improvement that was achieved using fully unsupervised learning, where for example the WER for ASR improved by 2.63% without any textual or audio data, and similar improvements occurred for all 4 tasks.

This architecture addresses one of the most prominent issues in multimodal learning today, the lack of availability of multimodal paired data, but the self-supervised and unsupervised approach could be a very useful tool for future works.

5 Challenges and Future Work

Multimodal learning is an open and very promising area of research. Results have shown that multimodal learning can be used to improve upon existing single modality tasks 2.1 and also enable new, unique capabilities not currently possible from a unimodal model 2.2. However, there are still some limitations and areas of improvement.

Data availability is an open problem with multimodal learning. There is currently a lack of abundance of large, paired datasets which can be used for multimodal architectures [21], which can limit the extent and utility of multimodal approaches. However, self-supervised approaches based on unpaired data [8] can be a useful tool for mitigating this problem. There are also a number of open efforts to create larger, paired datasets, such as MUGEN, a large video-audio-text dataset [11] or the Lip Reading in the Wild Dataset [5].

The results of outputs from unimodal deep learning architecture are sometimes difficult to interpret and explain, and issues of ethics and fairness of predictions have been raised in recent years [18]. Multimodal deep learning approaches can further exacerbate these issues and make results even harder to interpret and ensure fairness in results [21].

In recent years, advances in GPU acceleration have enabled very deep unimodal networks with wide applicability [19]. With the introduction of multimodal inputs and learning, the computational and memory requirements for equivalently deep networks will increase to be able to efficiently process a multiplicatively greater amount of data with a similar depth [21]. This can make real-time inference difficult, particularly on portable devices [23].

These are just some of the major challenges present in the field of multimodal learning. Some other challenges including finding the best practices or a general framework for fusion methods, as we discussed in Section 3, dealing with overfitting and varying learning rates for different input modalities, and reducing noise in multimodal input data [21].

6 Conclusion

This survey paper focused on multimodal learning involving speech. Two main themes are presented based on our findings: the performance improvement provided by utilizing multimodal learning for ASR, and new tasks/processing capabilities achievable by multimodal models. In addition, we provided an overview of various styles of multimodal fusion and how the works surveyed implemented multimodal fusion in their model architectures. Finally, we discussed the multimodal chain architecture, an alternate style of utilizing multimodal inputs. The simultaneous processing of and reasoning about multiple input modalities, which is central to multimodal learning, can provide context about a related group of data that is unachievable by unimodal processing. As a result of our survey, we believe that multimodal learning has strong potential to improve the performance of learning tasks involving speech and should continue to be studied.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

[2] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report, 2023.

[3] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: a language modeling approach to audio generation, 2023.

[4] William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals. Listen, attend and spell, 2015.

[5] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, 2016.

[6] Michael J. Crosse, Giovanni M. Di Liberto, and Edmund C. Lalor. Eye can hear clearly now: Inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. *Journal of Neuroscience*, 36(38):9888–9895, 2016.

[7] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context, 2019.

[8] Johanes Effendi, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Listening while speaking and visualizing: Improving asr through multimodal chain. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 471–478, 2019.

[9] Luis Espinosa-Anke, Thierry Declerck, Dagmar Gromann, Armand Vilalta, Dario Garcia-Gasulla, Ferran Parés, Eduard Ayguadé, Jesus Labarta, E. Ulises Moya-Sánchez, Ulises Cortés, Dagmar Gromann, Luis Espinosa Anke, and Thierry Declerck. Studying the impact of the full-network embedding on multimodal pipelines. *Semant. Web*, 10(5):909–923, jan 2019.

[10] Ivan Grishchenko, Artsiom Ablavatski, Yury Kartynnik, Karthik Raveendran, and Matthias Grundmann. Attention mesh: High-fidelity face mesh prediction in real-time. *CoRR*, abs/2006.10962, 2020.

[11] Thomas Hayes, Songyang Zhang, Xi Yin, Guan Pang, Sasha Sheng, Harry Yang, Songwei Ge, Qiyuan Hu, and Devi Parikh. Mugen: A playground for video-audio-text multimodal understanding and generation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 431–449, Cham, 2022. Springer Nature Switzerland.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[13] Suyoun Kim, Takaaki Hori, and Shinji Watanabe. Joint ctc-attention based end-to-end speech recognition using multi-task learning, 2017.

[14] Michelle A. Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. *CoRR*, abs/1810.10191, 2018.

[15] Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7001–7011, 2021.

[16] Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *SciPy*, 2015.

[17] HARRY MCGURK and JOHN MACDONALD. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.

[18] Chuizheng Meng, Loc Trinh, Nan Xu, and Yan Liu. Mimic-if: Interpretability and fairness evaluation of deep learning models on mimic-iv dataset, 2021.

[19] Mohit Pandey, Michael Fernandez, Francesco Gentile, Olexandr Isayev, Alexander Tropsha, Abraham C Stern, and Artem Cherkasov. The transformational role of GPU computing and deep learning in drug discovery. *Nat. Mach. Intell.*, 4(3):211–221, March 2022.

[20] Georgios Paraskevopoulos, Srinivas Parthasarathy, Aparna Khare, and Shiva Sundaram. Multiresolution and multimodal speech recognition with transformers, 2020.

[21] Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion*, 81:203–239, 2022.

[22] Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. Audiopalm: A large language model that can speak and listen, 2023.

[23] Dmitry Ryumin, Denis Ivanko, and Elena Ryumina. Audio-visual speech and gesture recognition by sensors of mobile devices. *Sensors*, 23(4), 2023.

[24] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: A large-scale dataset for multimodal language understanding, 2018.

[25] Bjorn Schuller, Florian Metze, Stefan Steidl, Anton Batliner, Florian Eyben, and Tim Polzehl. Late fusion of individual engines for improved recognition of negative emotion in speech - learning vs. democratic vote. 3 2010.

[26] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Fine-grained audio-visual joint representations for multimodal large language models, 2023.

[27] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Multi-scale alignment and contextual history for attention mechanism in sequence-to-sequence model, 2018.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[29] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR.